Efficient unsupervised clustering for spatial bird population analysis along the Loire river

Aurore Payen[1], Ludovic Journaux[1,2], Clément Delion[1], Lucile Sautot[1,3], Bruno Faivre[3] *

1- AgroSup Dijon
26 Boulevard Docteur Petitjean, 21000 Dijon

2- Université de Bourgogne - LE2I
Avenue Alain Savary 21000 Dijon - France

3- Université de Bourgogne - Biogéosciences
6 Boulevard Gabriel 21000 Dijon - France

Abstract. This paper focuses on application and comparison of Non Linear Dimensionality Reduction (NLDR) methods on natural high dimensional bird communities dataset along the Loire River (France). In this context, biologists usually use the well-known PCA in order to explain the upstream-downstream gradient.Unfortunately this method was unsuccessful on this kind of nonlinear dataset.The goal of this paper is to compare recent NLDR methods coupled with different data transformations in order to find out the best approach. Results show that Multiscale Jensen-Shannon Embedding (Ms JSE) outperform all over methods in this context.

1 Introduction

The longitudinal distribution pattern of organisms along rivers is a major research topic in ecology, initiates for invertebrates by Illies, for fishes by Huet and on bird population by Frochot & *al.*[3]. In these works, ornithologists want to analyze spatiotemporal distribution of bird communities in order to study river zonation by detecting ecological discontinuities due to geomorphology of landscapes (discontinuities resulting from birds species assemblage). Biologists usually use Principal Component Analysis (PCA) in order to explain the longitudinal distribution pattern and find discontinuities along the upstream-downstream gradient of the river. Unfortunately, despite the robust reputation of PCA analysis on real-life data, it appears that this method is not able, with 19.82% of contribution on the two first factorial axis, to account these discontinuities which are nevertheless forebode by specialists. So, PCA shows a strong limitation in this case. To overcome this problem, it is interesting to use Non Linear Dimensionality Reduction (NLDR) methods to transform high-dimensional data into a meaningful low dimension representation. Numerous studies have aimed to compare NLDR algorithms, usually using synthetic data such as a swissroll [6], but less for natural data. So, this paper explore and compare recent NLDR methods

*Data acquisition received financial support from the FEDER Loire, Etablissement Public Loire, DREAL de Bassin Centre (Etude des oiseaux nicheurs de la Loire et de l'Allier sur l'ensemble de leurs cours) to BF, and from the Région Bourgogne (PARI, Projet Agrale 5) to BF. Data analysis received support from the French Ministry of Agriculture (Bourse FCPR to LS).

with different data transformations in order to find out the best approach in this ecological context. This paper is organized as follows. Section 2 presents the real-life dataset used on this context, the data transformations and an NLDR methods overview with a comparison based on a quality assessment. Section 3 presents and discusses experimental results. Section 4 draws the conclusions.

2 Materials and methods

2.1 Real-life dataset and transformation

Our dataset comes from the census birds STORI[1] program for nesting birds along the Loire River (France) [3]. STORI aims to observe spatiotemporal changes into bird populations along rivers.198 census points were defined along the Loire. At each point birds are identified with the PAI (Punctual Abundance Index) method [3] during four census campaigns. Bird abundances were described by a semi-quantitative abundance index. One of the main objective is to study global/local factors that explain bird abundances changes. Finally, we consider 140 birds species along the 198 census points. In practice, ornithologists capped PAI to 5 even if there is more than 5 couples of birds. Unfortunately, this approximation is not relevant because variables are expected to have a Gaussian distribution. Distribution of the variables are centered on the value of two couples. The number of couples should decrease, passed the center of the distribution. So the number of 4 couples should be superior than the number of 5 couples. 23 species among 140 have more PAI of 5 than 4, which means the census is over estimated. So in order to preserve or improve the gaussian distribution and corrects the dat we propose to apply transformation that correct skewness and kurtosis of distribution. Skewness and kurtosis are two measures that are zero when a variable is gaussian. We retain three transformations to correct Skewness and kurtosis: the square, the square root and the Anscombe transformation (AT). To select data transformation, we refere in section 3.1 to a quality criterion that check the impacts of the transformation: the transformation that gets the best quality criterion result has the best trade-off between Gaussian distribution and underestimate of bird couples.

2.2 Overview of different methods of dimensional reduction

NLDR methods can be classified according to different characteristics: **Scale analysis (local/global/multiscale)**. This reflects the kind of properties the transformation does preserve. **Distance metrics/similarity**. This shows the distance used to estimate if two data points are close. We retained 7 NLDR methods completed with to 2 linear methods: the Classical Multidimensional Scaling (CMDS) and the Non-metric Multidimensional Scaling (NMDS) [6]. ***Non linear Mapping (NLM)*** (Sammon's mapping) [6] tries to preserve the neigborhood topology of data by preserving distances between points according to the following stress function:

[1] Temporal Monitoring of Nesting Birds in River Valley

$$J_{NLM} = \frac{1}{\sum_{i,j=1}^{n} d_{i,j}^{m}} \left(\sum_{i,j=1}^{n} \frac{(d_{i,j}^{m} - d_{i,j}^{p})^2}{d_{i,j}^{m}} \right)$$

With d_{ij}^{m} and d_{ij}^{p} are the distances between points i^{th} and j^{th}, in $\mathbb{R}^m$and $\mathbb{R}^p$. ***Curvilinear Component Analysis (CCA)***[6] is an evolution of NLM. Instead of the optimization of a reconstruction error, CCA aims to preserve the distance matrix while projecting data onto $\mathbb{R}^p$ dimension, giving priority to low distances. The use of similarities in NLDR is recent[4].These approaches are based on sparse matrices of similarities defined in $\mathbb{R}^m$, such as in ***Stochastic Neighbor Embedding (SNE)*** [1] where distances are converted into probabilities which represent similarities. SNE aims to preserve similarities in $\mathbb{R}^m$ and $\mathbb{R}^p$. In this context ***t-distributed Stochastic Neighbor Embedding (t-SNE)***[2] and ***Neighbor Retrieval Visualizer (NeRV)***[8] are SNE evolutions. The first is based on Student t-distribution to calculate similarities while a Gaussian distribution is used in SNE. Both SNE and t-SNE try to reduce the Kullback–Leibler divergence (KLD) as a cost function. The second uses two dual KLD, that are more precise than a single KLD. Coming from this KLD approach, different refinement have been proposed. First by replacing KLD by Jensen-Shannon divergence in the ***Jensen-Shannon Embedding (JSE)***[5]. Secondly, to overcome one of the major drawback fixed size of neigborhood, [4] proposed in ***Multiscale Jensen-Shannon Embedding (Ms JSE)*** to take into account different sizes of neigborhood, thanks to a log scale.

2.3 Objective comparison based on quality assessment

Several quantitative evaluation measures for NLDR have been proposed including techniques which rely on neighborhood ranking. We based our quality criterion on the intrusion/extrusion diagram proposed by Lee & Verleysen. Fore more details on quality assessment see on[7]. This criterion is the Area Under Curve (AUC), a scale-independent criterion got by calculating the area under the curve of R_{NX} function, which gives the percentage of improvement of neighborhood preservation compared to a random projection, depending on the size of the neighborhood.

3 Results and discussions

3.1 The choice of a transformation

The table 2 shows AUC results for each transformation in order to measure their impacts on data. Results show for every NLDR methods that the square root gets better results than the data without transformation or AT. Moreover, the best result is obtain with MS JSE with 57,2%. Finally, we select the square root which correspond to the best trade-off between correcting the underestimate of bird couples and the Gaussian distribution.

Without transformation	Square root	Transformation of Anscombe
21.5 CMDS 30.7 NMDS 26.1 NLM 29.5 CCA 33.8 SNE 42.8 t-SNE 38.0 NeRV 45.9 JSE 49.4 Ms. JSE	27.3 CMDS 34.3 NMDS 27.5 NLM 31.2 CCA 43.6 SNE 50.0 t-SNE 46.8 NeRV 52.8 JSE 57.2 Ms. JSE	26.6 CMDS 33.0 NMDS 28.0 NLM 31.8 CCA 43.0 SNE 46.2 t-SNE 45.6 NeRV 50.9 JSE 55.1 Ms. JSE

Table 1: AUC results for data transformation

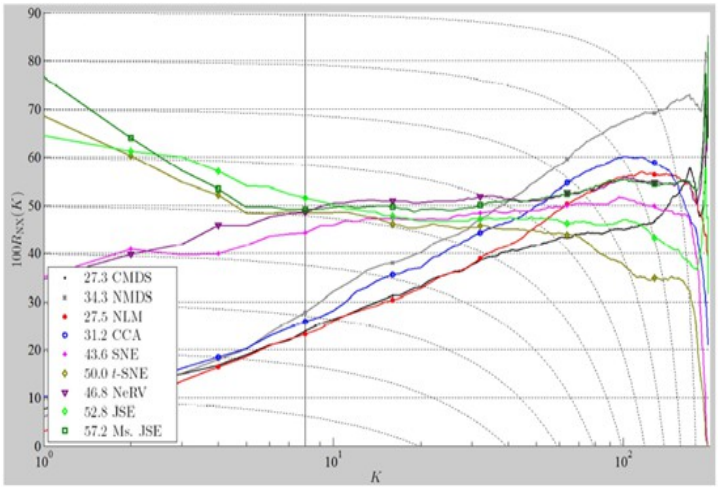

Figure 1: Quality curves of R_{NX} function on the square root of the data

3.2 The choice of the more efficient NLDR method

Figure 1 presents the quality curves coming from the differents methods. At lower scale, Ms JSE reaches 75% of improvement of neighborhood conservation compared to a random projection, and clearly outperforms other methods. At high scale, Ms JSE gets more than 70%, but is in the average compared to other methods. NMDS seems to be the best method for high neighborhoods (85% of iprovement for100 neighbors). NMDS R_{NX} function reaches faster than the other curves high results for high neighborhoods. But all the methods present good results for high neighborhoods, while Ms JSE, t-SNE and JSE are the only ones to get good results at small neighborhoods. For instance, NMDS result is very low, about 10%.Ms JSE has better results than t-SNE and JSE at large neighborhoods, that's why Ms JSE has better AUC result than any other method, with 57,2%. Finally, Ms JSE seems to be the best trade-off for this ecological application.

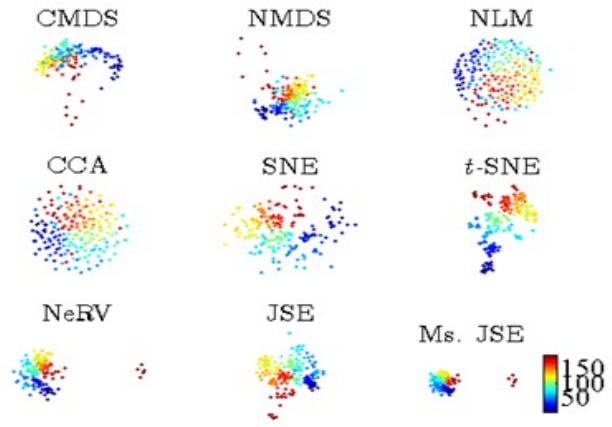

Figure 2: Projection of the square root of the data, with all the methods (blue: upstream of Loire river, red brown: downstream)

3.3 Resulting projection of the data with NLDR methods

Census points are projected with methods (Figure 2). The projection comparison shows that NLDR based on similarity outperform over methods in clustering context. Focusing on Ms JSE (Figure 2). we can see that global/local organization of data is respected. At global scale we observe the upstream-downstream gradient relationship between census points. At local scale, Ms JSE is able to give an efficient clustering, grouping the data which have the same characteristics in terms of birds species assemblage. In the context of river zonation, each cluster represents a different birds species assemblage depending on environmental features and each distance between clusters represent an ecological discontinuity which is not detectable with linear approach. Moreover, Ms JSE has the characteristic of preserving outliers. In fact, the five last census points are distinct from the others: they are the five last census points, located in the downstream of the Loire River (next to the Atlantic Ocean). These census points are indeed very different from the others considering their bird population. This distinguish with the other census points isn't clear with local NLDR methods, such as NLM, CCA, SNE and t-SNE.

4 Conclusion

This paper explores and compares recent NLDR methods with data transformations in order to find the best method on a real-life ornithological application. Results highlight that Ms JSE with square root transformation is the most efficient method. The global organization of census points reveals the upstream-downstream gradient and the local clustering highlights discontinuities. These results outperform traditional PCA in this context. In order to generalise results, more tests on another nonlinear natural dataset should be made to validate if Ms JSE confirms its ability to make efficient projections in ecological context.

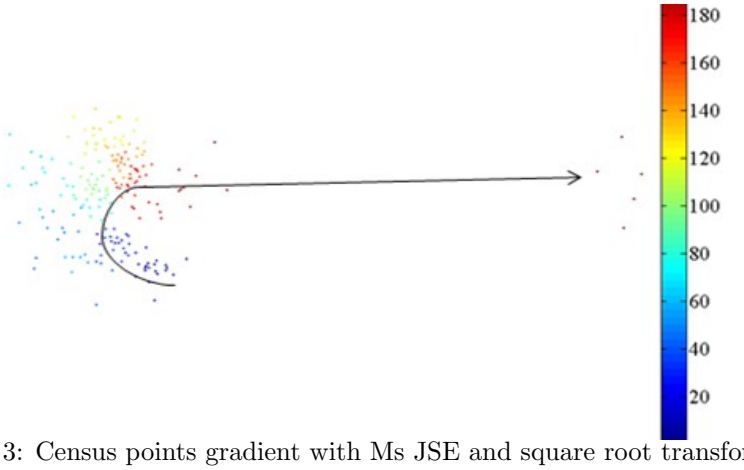

Figure 3: Census points gradient with Ms JSE and square root transformation

References

[1] Kerstin Bunte, Sven Haase, and Michael Biehl et al. Stochastic neighbor embedding (sne) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

[2] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[3] B. Frochot, M.C. Eybert, L. Journaux, J. Roché, and B. Faivre. Nesting birds assemblages along the river loire: result from a 12 years-study. *Alauda*, 71(2):179–190, 2003.

[4] John A. Lee, Diego H. Peluffo-Ordonez, and Michel Verleysen. Multiscale stochastic neighbor embedding: Towards parameter-free dimensionality reduction. In *Proceedings of 22st European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning (ESANN)*, 2014.

[5] John A Lee, Emilie Renard, Guillaume Bernard, Pierre Dupont, and Michel Verleysen. Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[6] John Aldo Lee and Michel Verleysen. *Nonlinear Dimensional Reduction*. Information Science and Statistics. Springer, 2007.

[7] John Aldo Lee and Michel Verleysen. Quality assessment of dimensionality reduction: rank-based criteria. *Neurocomputing*, 72:1431–1443, 2009.

[8] Jarkko Venna, Jaakko Peltonen, and Kristian Nybo et al. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.